þÿ S I A K A G a m e f o r F o r e i g n L a n g u a g e P r o n u n c i a t

Karhila, Reima

International Speech Communications Association
2017

Karhila , R , Ylinen , S P , Enarvi , S , Palomäki , K , Nikulin , A , Rantula , O , Viitanen , V , Dhinakaran , K , Smolander , A-R M , Kallio , H H , Uther , M , Junttila , K P M & Kurimo , M þÿ 2 0 1 7 , S I A K A G a m e f o r F o r e i g n L a n g u a g e P r o n u n c i a t i o n L e a r n i n g . i INTERSPEECH 2017 : Interspeech: Annual Conference of the International Speech Communication Association . vol. 2017 , Interspeech: Annual Conference of the International Speech Communication Association , International Speech Communications Association , Stockholm , pp. 3429-3430 , Interspeech 2017 , Stockholm , Sweden , 20/08/2017 . < https://www.isca-speech.org/archive/Interspeech_2017/pdfs/2046.PDF >

http://hdl.handle.net/10138/313214

publishedVersion

SIAK – A Game for Foreign Language Pronunciation Learning

Sari Ylinen[2], *Reima Karhila*[1], *Seppo Enarvi*[1], *Kalle Palomäki*[1], *Aleksander Nikulin*[1], *Olli Rantula*[1], *Vertti Viitanen, Krupakar Dhinakaran*[1], *Anna-Riikka Smolander*[2], *Heini Kallio*[2], *Maria Uther*[3], *Katja Junttila*[2], *Mikko Kurimo*[1]

[1]Aalto University, Finland
[2]University of Helsinki, Finland
[3]University of Winchester, England
sari.ylinen@helsinki.fi, mikko.kurimo@aalto.fi

Abstract

We introduce a digital game for children's foreign-language learning that uses automatic speech recognition (ASR) for evaluating children's utterances. Our rst prototype focuses on the learning of English words and their pronunciation. The game connects to a network server, which handles the recognition and pronunciation grading of children's foreign-language speech. The server is reusable for different applications. Given suitable acoustic models, it can be used for grading pronunciations in any language.

Index Terms: pronunciation grading, language learning, phonetics, speech analysis

1. Introduction

Language and communication skills play an important role in our society. The benefits of early foreign language learning are acknowledged, but often the learning of foreign languages starts in school so late that the optimal period for language learning has already passed [1]. If foreign-language learning was based on speech rather than written textbooks, it could be started before school-age. If the foreign language was learned from speech similarly to native language, brain representations for foreign speech sounds and words would likely end up being more native-like. Speech-based learning applications and games have potential for this kind of language teaching. However, to be interactive, speech-based applications need an effective speech interface, which requires advancements in speech technology. The problem is that the state-of-the-art speech technology is not adaptive enough to recognize child language learner's speech.

We have designed a computer game called Say it again, kid! (SIAK) that uses ASR for the assessment of children's speech in a foreign language. The aim of the game is to teach foreign words and their pronunciation by encouraging children to listen and produce speech. Listening and speaking is expected to gradually promote the establishment of phonetic categories and word representations in the brain. Both processes are dependent on the neural connections between auditory brain areas in the temporal lobe and motor areas in the frontal lobe. By eliciting speech in a foreign language, we aim to establish auditory-motor brain representations that resemble those of native speakers. Importantly, category learning utilizes subcortical areas that are responsive to feedback [2]. We therefore activate these brain areas by providing instantaneous feedback in the form of a score computed using speech recognition technology after each utterance.

SIAK is implemented as a board game. In the current version, the target group is Finnish children with little or no experience in English. The game presents the player a board, where the player can move. There are many boards in the game world. After collecting enough points and working their way through the board, the players can move to the next board and jump between unlocked boards (see Figure 1). A board contains a number of cards that the player can open. Each card introduces a new English word. Later in the game, cards may contain sentences consisting of a few words. Upon opening a card, the player hears the word in Finnish and in English (produced by different native English speakers) and sees a related picture. The child's task is to imitate the word aloud.

Figure 1: *SIAK game records player's utterances using a headset. The game allows the player to move between unlocked boards.*

Another component of the system is a network server,

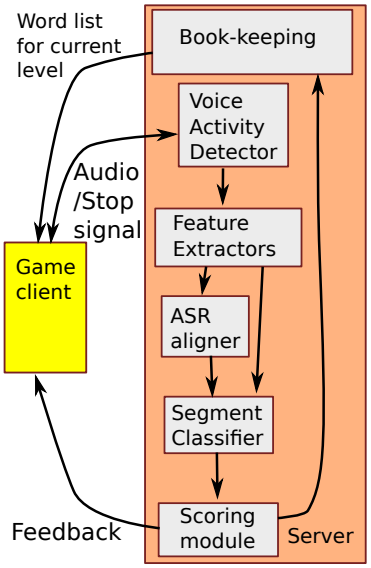

Figure 2: *Block diagram of scoring system.*

which is responsbile for grading the words uttered by the player. The game sends a recorded word to the server, and the server returns a numerical score. Then the child's own and native English speaker's utterances are played again for comparison, and the player receives one to five points based on the utterance score. Increasing scores open new routes on the game board for exploration. In order to keep the game motivating, the players are expected to perform various gimmicks before they can finish a board. In addition to imitation, children are encouraged to test their word learning in test cards where they should recall and say the English word without the model pronunciation.

2. School trials

We are currently starting trials for evaluating the game in classrooms. At this stage, participants are 9-year-old Finnish children who have recently started learning English, but later also younger children with no foreign-language experience will be of interest. During gameplay, each player has an Android tablet or a Windows PC and a headset. Children will play approximately 15 minutes per day, 5 days a week for six weeks. During this period, we will collect a database of children's recorded utterances, part of which will be analyzed acoustically. We will also follow up players' progress and different paths of learning from assessment scores provided by the grading system.

In addition, to demonstrate plastic changes in the brain induced by gaming, we will measure childrens brain responses to foreign words with electroencephalography (EEG) before and after the gaming period. Brain responses will also be used to investigate how the accuracy of feedback given by the ASR system affects learning. Specifically, we will compare learning in child groups that receive either accurate realistic feedback or feedback that is not dependent on children's performance.

3. System architecture

The game has been developed using Unity game engine. It can be exported to over 25 platforms with relative ease. Currently we have created Windows and Android builds. When the game is started, it asks the player to sign in using a username and password. The player is identified in order to keep track of the progress of each player and possible in the future adapt the models to each player's voice.

Another component of the game is the network server. It is used for grading speech and storing the game state between sessions. When the player moves to a card on the game board, native English pornunciation of the word in question is played through the headset, the player repeats the word, and the player's pronunciation is recorded. The recording is streamed to the server, and the server starts to analyze the audio as soon as it receives the first packet.

The server contains two components: a speech aligner and a phoneme classifier. First forced-alignment is performed using Aalto ASR [3], a GMM-HMM speech recognizer. Decoding is not needed, because the target word is known in advance. As a result of the alignment, the system gets a mapping of time frames to phonemes. The second part of the evaluation consists of classifying each frame to a phoneme using a DNN classifier. The pronunciation score is computed by comparing the phonemes obtained using the DNN classifier to the phonemes obtained using forced-alignment.

When all audio packets are sent to the server, it computes the score and sends it back to the game.

4. References

[1] J. Johnson and E. Newport, "Critical period effects in second-language learning: The influence of maturational state on the acquisition of english as a second language," *Cognitive Psychology*, vol. 21, pp. 60–99, 1989.

[2] H.-G. Yi, W. T. Maddox, J. A. Mumford, and B. Chandrasekaran, "The role of corticostriatal systems in speech category learning," *Cerebral Cortex*, vol. 26, no. 4, pp. 1409–1420, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4785939/

[3] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 4, pp. 724–732, 2009. [Online]. Available: http://dx.doi.org/10.1109/TASL.2008.2012323